University of BRISTOL

Discussion of
Analysis of Forensic DNA Mixtures with Artefacts

by Cowell, Graversen, Lauritzen and Mortera

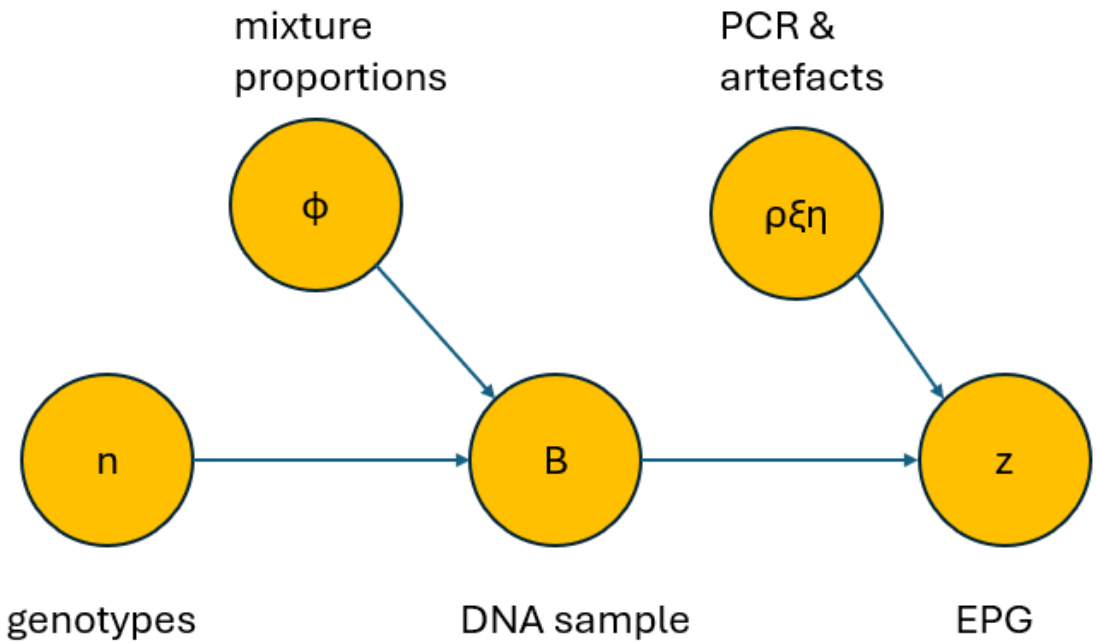
Peter Green

RSS Journal Webinar, August 2025

CGLM model

CGLM provide a joint probability model connecting

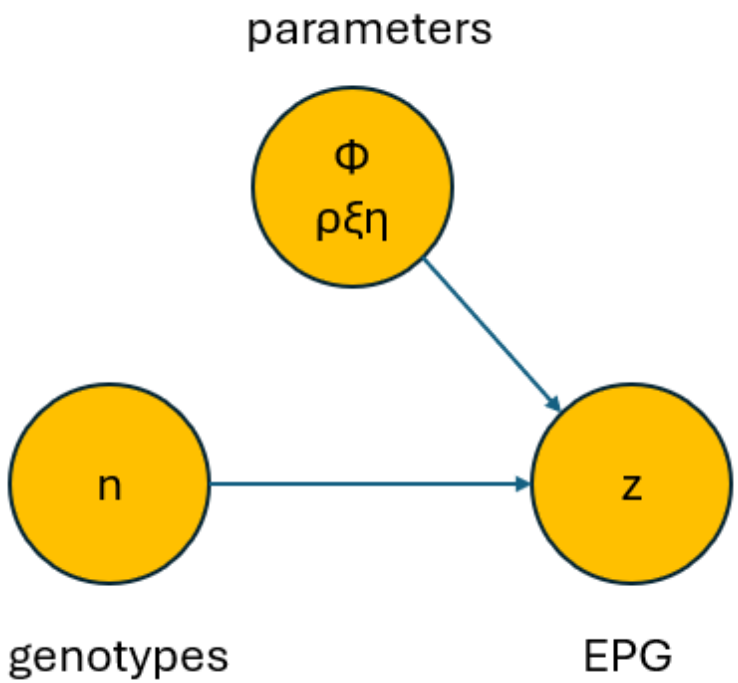
- EPG peak heights z
- genotypes of contributors n
- hypotheses $\mathcal{H}$ about contributors



CGLM recap

It is natural to assume

$$p(z, n|\mathcal{H}) = p(n|\mathcal{H}) \times p(z|n)$$



$$p(z, n|\mathcal{H}) = p(n|\mathcal{H}) \times p(z|n)$$

The key (and innovative!) ingredients

- gamma peak height model for $p(z|n)$
- captures (most) artefacts
- Bayes net representation for $p(n|\mathcal{H})$: continuous-valued z are fixed, so $p(n|\mathcal{H}, z)$ is still a BN for n
- 'allele count' representation for n – Markov serial dependence along allele values
- DNAmixtures R package

Paternity testing

A common set-up:

- a woman claims that a certain individual (the 'putative father') is the father of her child
- mother, child and putative father are genotyped (using STR markers)
- we wish to assess the evidence for paternity

The plaintiff's hypothesis is $\mathcal{H}_p$: the putative father is the true father; the alternative is $\mathcal{H}_0$: someone else is the father – to be precise, the father is a randomly chosen person (male) in the population (which population?)

But what if the putative father's genotype is not available? All we have is a biological sample that appears to be a mixture of material from several individuals, possibly including the putative father.

Paternity testing

A common set-up:

- a woman claims that a certain individual (the 'putative father') is the father of her child
- mother, child and putative father are genotyped (using STR markers)
- we wish to assess the evidence for paternity

The plaintiff's hypothesis is $\mathcal{H}_p$: the putative father is the true father; the alternative is $\mathcal{H}_0$: someone else is the father – to be precise, the father is a randomly chosen person (male) in the population (which population?)

But what if the putative father's genotype is not available? All we have is a biological sample that appears to be a mixture of material from several individuals, possibly including the putative father.

Paternity testing

A common set-up:

- a woman claims that a certain individual (the 'putative father') is the father of her child
- mother, child and putative father are genotyped (using STR markers)
- we wish to assess the evidence for paternity

The plaintiff's hypothesis is $\mathcal{H}_p$: the putative father is the true father; the alternative is $\mathcal{H}_0$: someone else is the father – to be precise, the father is a randomly chosen person (male) in the population (which population?)

But what if the putative father's genotype is not available? All we have is a biological sample that appears to be a mixture of material from several individuals, possibly including the putative father.

Relatedness, relationships and DNA mixtures

... based on joint work with Julia Mortera

In reality, the genotypes of actors (mixture contributors and other typed individuals) are not independent. We distinguish between

- (ambient) relatedness: there is always a degree of kinship among individuals in sub-populations: island model, θ correction, etc
- (specific, close-family) relationships: under some hypothesised scenarios, some actors are close relatives (brothers in crime, paternity cases, etc.)

and in the latter case we may require analysis

- about such relationships (e.g. paternity)
- allowing for such relationships (e.g. was suspect's brother also at the scene?)

Relatedness, relationships and DNA mixtures

... based on joint work with Julia Mortera

In reality, the genotypes of actors (mixture contributors and other typed individuals) are not independent. We distinguish between

- (ambient) relatedness: there is always a degree of kinship among individuals in sub-populations: island model, θ correction, etc
- (specific, close-family) relationships: under some hypothesised scenarios, some actors are close relatives (brothers in crime, paternity cases, etc.)

and in the latter case we may require analysis

- about such relationships (e.g. paternity)
- allowing for such relationships (e.g. was suspect's brother also at the scene?)

CGLM recap

CGLM provide a joint probability model connecting

- EPG peak heights z
- genotypes of contributors n
- hypotheses $\mathcal{H}$ about contributors

$$p(z, n|\mathcal{H}) = p(n|\mathcal{H}) \times p(z|n)$$

In the standard case, for $p(n|\mathcal{H})$ we would assume that the genotypes of the k contributors are k independent draws from the 'gene pool'. Here, we will show how to replace this so that $p(n|\mathcal{H})$ models the hypothesised relationships (among contributors, or between contributors and typed individuals).

CGLM recap

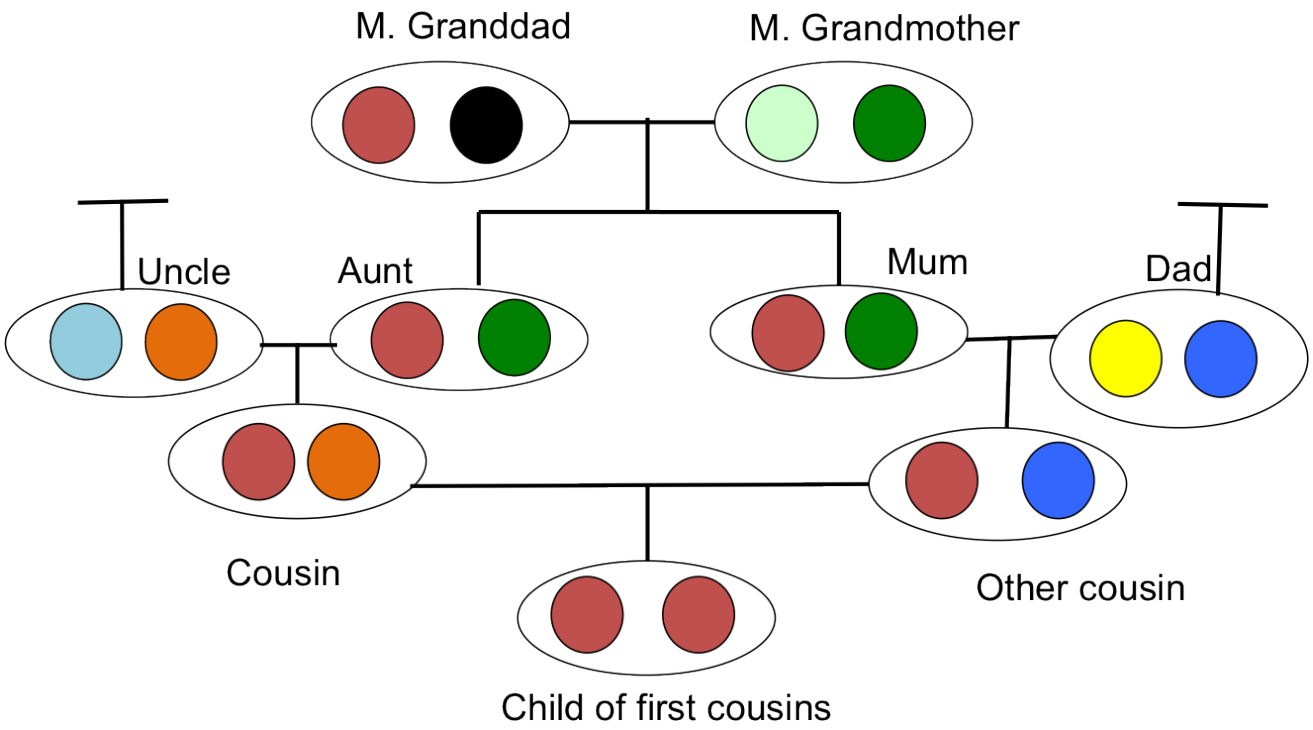
CGLM provide a joint probability model connecting

- EPG peak heights z
- genotypes of contributors n
- hypotheses $\mathcal{H}$ about contributors

$$p(z, n|\mathcal{H}) = p(n|\mathcal{H}) \times p(z|n)$$

In the standard case, for $p(n|\mathcal{H})$ we would assume that the genotypes of the k contributors are k independent draws from the 'gene pool'. Here, we will show how to replace this so that $p(n|\mathcal{H})$ models the hypothesised relationships (among contributors, or between contributors and typed individuals).

Identity by descent



(none of this is specific to CGLM, STR markers, or forensic genetics at all)

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships $\equiv$ patterns of IBD

IBD – Identity by descent.

We consider here only autosomal STR markers, one marker (locus) at a time, under Hardy-Weinberg equilibrium.

Consider a parent and child: the joint distribution of their genotypes can be precisely described thus: the parent's genotype is (a, b) and the child's is (a, c), where a, b, c are independently drawn from a distribution q over alleles (the allele frequency distribution for this population and marker). The a appears in both genotypes: the genes are *identical by descent*.

In the absence of inbreeding, *any* two relatives' genotypes can be written as (a, b), and (a, b), (a, c) or (c, d) with probabilities $\kappa_2, \kappa_1, \kappa_0$, where $a, b, c, d \overset{iid}{\sim} q$, where $\kappa_0 + \kappa_1 + \kappa_2 = 1$.
e.g. for parent & child, $\kappa_0 = \kappa_2 = 0, \kappa_1 = 1$
for siblings, $\kappa_0 = \kappa_2 = 0.25, \kappa_1 = 0.5$

Family relationships ≡ patterns of IBD

For two individuals, but allowing also for inbreeding, the situation was captured by Jacquard's (1974) condensed coefficients of descent $\Delta_1, \ldots, \Delta_9$.

The general case of any number of individuals has been formulated and analysed in seminal work by Elizabeth Thompson, using models for the meioses in the pedigree – the random selection of which gene each parent passes to their offspring.

We choose to represent relationships within a pedigree using the IBD pattern distribution.

Family relationships $\equiv$ patterns of IBD

For two individuals, but allowing also for inbreeding, the situation was captured by Jacquard's (1974) condensed coefficients of descent $\Delta_1, \ldots, \Delta_9$.

The general case of any number of individuals has been formulated and analysed in seminal work by Elizabeth Thompson, using models for the meioses in the pedigree – the random selection of which gene each parent passes to their offspring.

We choose to represent relationships within a pedigree using the IBD pattern distribution.

Family relationships $\equiv$ patterns of IBD

For two individuals, but allowing also for inbreeding, the situation was captured by Jacquard's (1974) condensed coefficients of descent $\Delta_1, \ldots, \Delta_9$.

The general case of any number of individuals has been formulated and analysed in seminal work by Elizabeth Thompson, using models for the meioses in the pedigree – the random selection of which gene each parent passes to their offspring.

We choose to represent relationships within a pedigree using the IBD pattern distribution.

IBD pattern distributions

IBD pattern distributions for a Father/Mother/Child triple.

(a) Distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
0.25	1	2	3	4	1	3
0.25	1	2	3	4	2	3
0.25	1	2	3	4	1	4
0.25	1	2	3	4	2	4

(b) Condensed form $\equiv$ Not distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
1	1	3	2	4	1	2

(c) Extending the family to include the paternal grandfather

pr	Fgt		Mgt		Cgt		GFgt	
0.5	1	3	2	4	1	2	1	5
0.5	1	3	2	4	1	2	3	5

IBD pattern distributions

IBD pattern distributions for a Father/Mother/Child triple.

(a) Distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
0.25	1	2	3	4	1	3
0.25	1	2	3	4	2	3
0.25	1	2	3	4	1	4
0.25	1	2	3	4	2	4

(b) Condensed form $\equiv$ Not distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
1	1	3	2	4	1	2

(c) Extending the family to include the paternal grandfather

pr	Fgt		Mgt		Cgt		GFgt	
0.5	1	3	2	4	1	2	1	5
0.5	1	3	2	4	1	2	3	5

IBD pattern distributions

IBD pattern distributions for a Father/Mother/Child triple.

(a) Distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
0.25	1	2	3	4	1	3
0.25	1	2	3	4	2	3
0.25	1	2	3	4	1	4
0.25	1	2	3	4	2	4

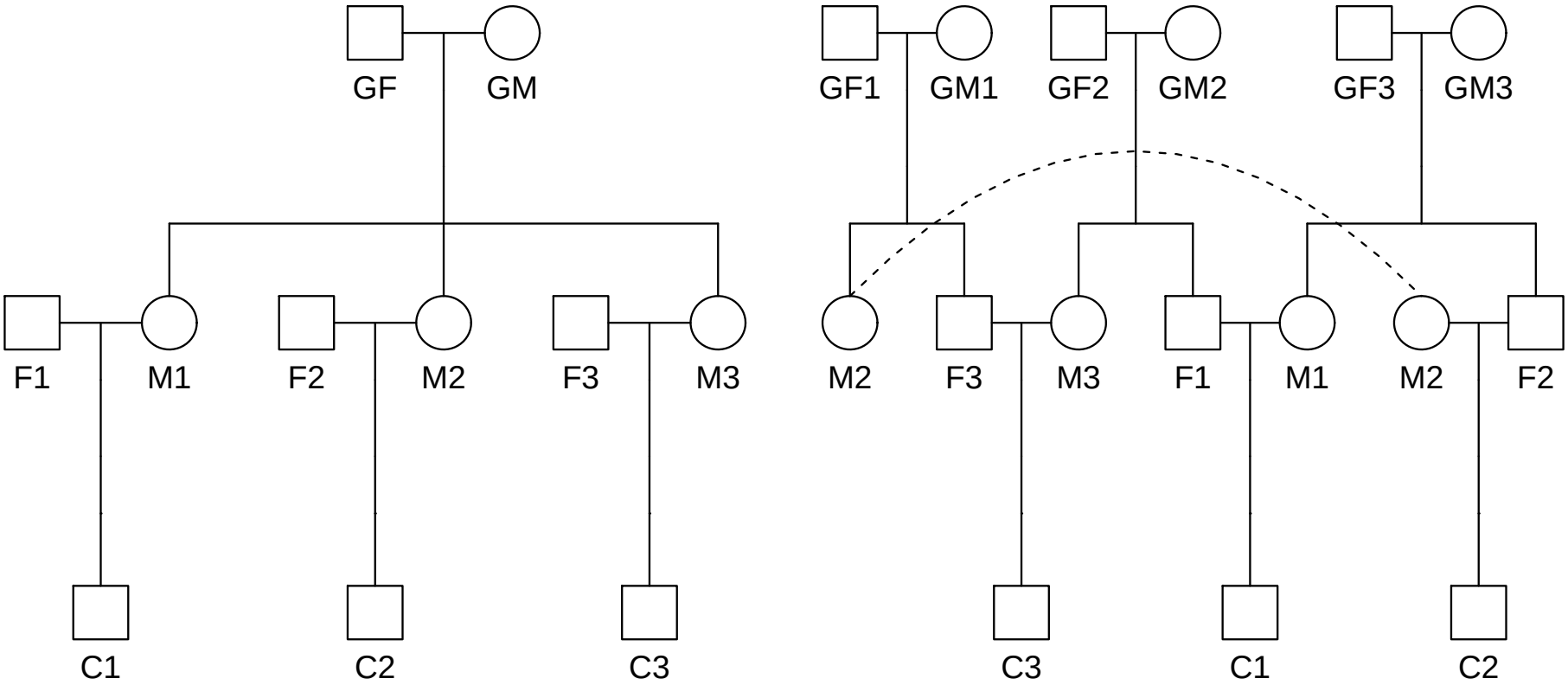
(b) Condensed form $\equiv$ Not distinguishing maternal and paternal genes

pr	Fgt		Mgt		Cgt	
1	1	3	2	4	1	2

(c) Extending the family to include the paternal grandfather

pr	Fgt		Mgt		Cgt		GFgt	
0.5	1	3	2	4	1	2	1	5
0.5	1	3	2	4	1	2	3	5

Pairwise relationships do not determine joint ones



Pairwise relationships do not determine joint ones

IBD pattern distributions for two scenarios of 3 pairwise cousins; (left) star, (right) cyclic arrangements.

pr	C1		C2		C3		pr	C1		C2		C3	
0.3750	1	2	3	4	5	6	0.421875	1	2	3	4	5	6
0.1875	1	2	1	3	4	5	0.140625	1	2	1	3	4	5
0.1875	1	2	3	4	1	5	0.140625	1	2	3	4	1	5
0.1875	1	2	3	4	3	5	0.140625	1	2	3	4	3	5
0.0625	1	2	1	3	1	4	0.046875	1	2	1	3	2	4
							0.046875	1	2	1	3	3	4
							0.046875	1	2	3	4	1	3
							0.015625	1	2	1	3	2	3

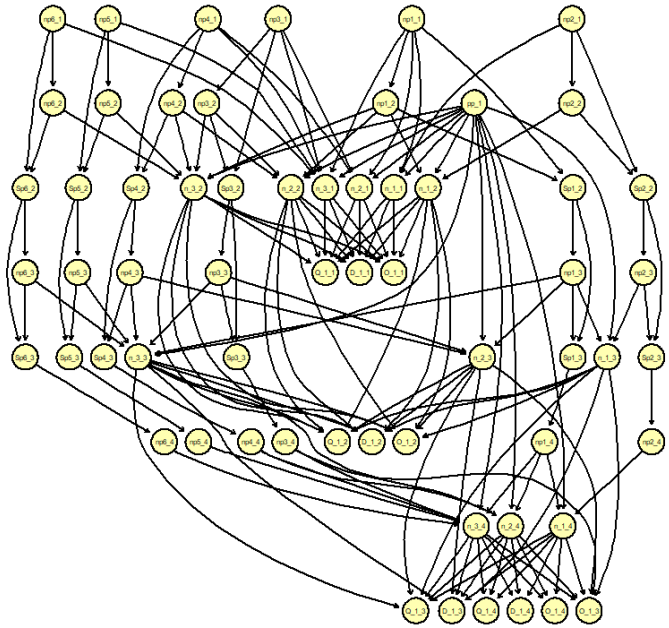
Modelling and computation in KinMix

In

$$p(z, n|\mathcal{H}) = p(n|\mathcal{H}) \times p(z|n)$$

we replace the Bayes net representing the standard $p(n|\mathcal{H})$ by one generated automatically from the IBD pattern distribution.

```
> as.IBD('3cousins-star')
     pr a   b   c
 0.0625 1 2 1 3 1 4
 0.1875 1 2 1 3 4 5
 0.1875 1 2 3 4 1 5
 0.1875 1 2 3 4 3 5
 0.3750 1 2 3 4 5 6
```



Software

Therese Graversen's R package DNAmixtures performs DNA mixture analysis using the CGLM model.

My R package KinMix augments DNAmixtures to implement the new methods presented here.

DNAmixtures and hence KinMix requires the commercial software Hugin (via the RHugin package).

DNAmixturesLite and KinMixLite are freeware versions of these, and are available on CRAN – with slight limits on functionality and problem size, and some loss of efficiency (with Therese Graversen).

https://petergreen5678.github.io/research/software/kinmix.html

Software

Therese Graversen's R package DNAmixtures performs DNA mixture analysis using the CGLM model.

My R package KinMix augments DNAmixtures to implement the new methods presented here.

DNAmixtures and hence KinMix requires the commercial software Hugin (via the RHugin package).

DNAmixturesLite and KinMixLite are freeware versions of these, and are available on CRAN – with slight limits on functionality and problem size, and some loss of efficiency (with Therese Graversen).

https://petergreen5678.github.io/research/software/kinmix.html

Software

Therese Graversen's R package DNAmixtures performs DNA mixture analysis using the CGLM model.

My R package KinMix augments DNAmixtures to implement the new methods presented here.

DNAmixtures and hence KinMix requires the commercial software Hugin (via the RHugin package).

DNAmixturesLite and KinMixLite are freeware versions of these, and are available on CRAN – with slight limits on functionality and problem size, and some loss of efficiency (with Therese Graversen).

https://petergreen5678.github.io/research/software/kinmix.html

Software

Therese Graversen's R package DNAmixtures performs DNA mixture analysis using the CGLM model.

My R package KinMix augments DNAmixtures to implement the new methods presented here.

DNAmixtures and hence KinMix requires the commercial software Hugin (via the RHugin package).

DNAmixturesLite and KinMixLite are freeware versions of these, and are available on CRAN – with slight limits on functionality and problem size, and some loss of efficiency (with Therese Graversen).

https://petergreen5678.github.io/research/software/kinmix.html

Software

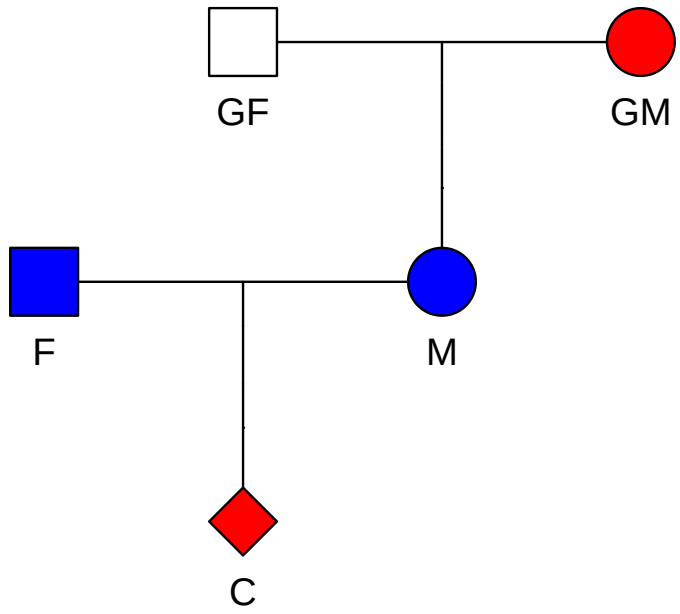
Therese Graversen's R package DNAmixtures performs DNA mixture analysis using the CGLM model.

My R package KinMix augments DNAmixtures to implement the new methods presented here.

DNAmixtures and hence KinMix requires the commercial software Hugin (via the RHugin package).

DNAmixturesLite and KinMixLite are freeware versions of these, and are available on CRAN – with slight limits on functionality and problem size, and some loss of efficiency (with Therese Graversen).

https://petergreen5678.github.io/research/software/kinmix.html




```
IBD<-as.IBD(matrix(c('M','GM','GF',
   'C','M','F'),2,3,byrow=TRUE),
  targets=c('F','M','C','GM'))

mix<-KinMix(list(epg),k=2,C=list(0.001),database=db,
 contribs=c('F','M'),typed.gts=list(C=Cgt,GM=GMgt),
 IBD=IBD,targets=c('F','M','C','GM'))
```

References

Cowell, R. G., Graversen, T., Lauritzen, S., and Mortera, J. (2015). Analysis of DNA mixtures with artefacts. *Journal of the Royal Statistical Society Series C* (with discussion), 64, 1–48 (my discussion p.41).

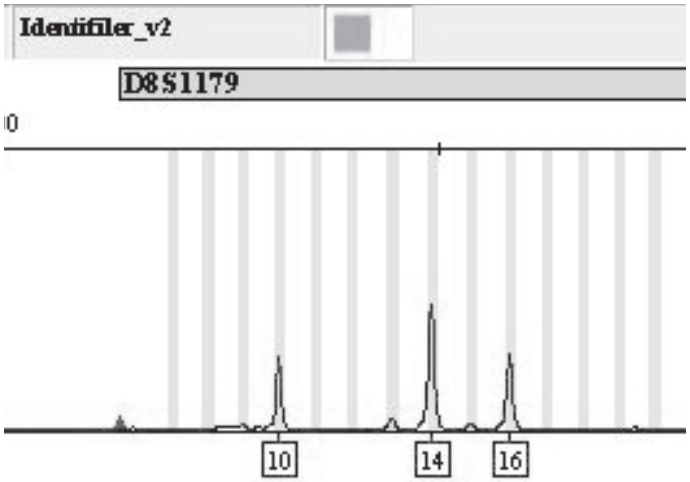
Green, P. J. and Mortera, J. (2009) Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Annals of Applied Statistics*, 3, 731-763. doi: 10.1214/09-AOAS235.
not mixtures: heterogeneous populations, uncertain allele frequencies, UAF=coancestry

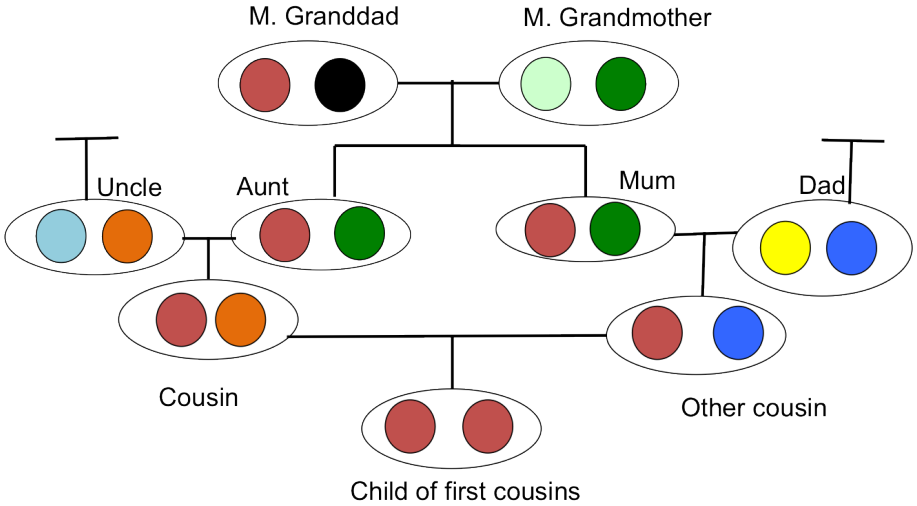
Green, P. J. and Mortera, J. (2017). Paternity testing and other inference about relationships from DNA mixtures. *Forensic Science International: Genetics*, 28, 128–37.
manipulating BNs to express relationships

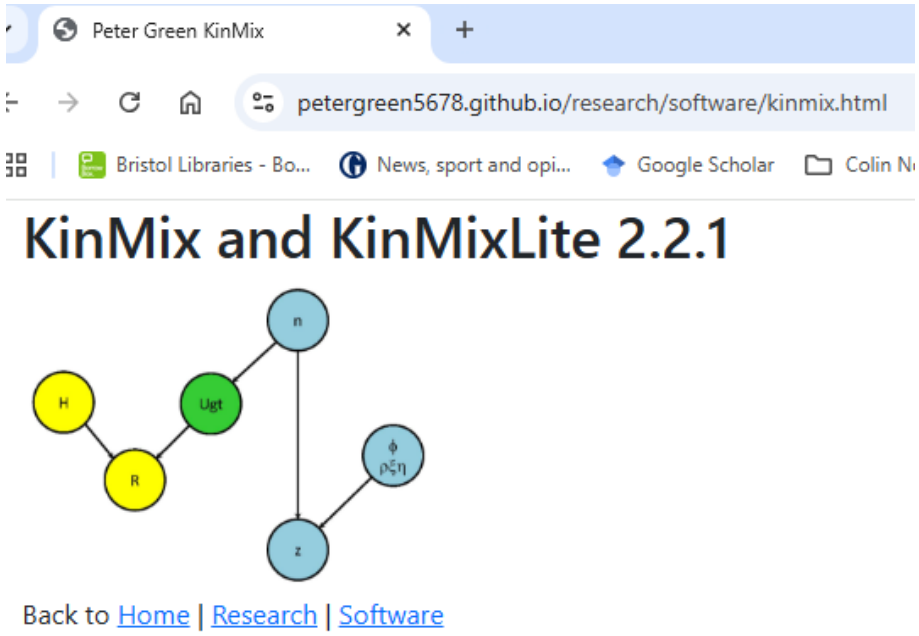
Green, P. J. and Mortera, J. (2021). Inference about complex relationships using peak height data from DNA mixtures. *Journal of the Royal Statistical Society Series C*, 70, 1049–1082.
general IBD patterns, model, algorithm, KinMix package

Green, P. J., Mortera, J., and Prieto, L. (2021). Casework applications of probabilistic genotyping methods for DNA mixtures that allow relationships between contributors. *Forensic Science International: Genetics*, 52, 102482. casework examples

https://petergreen5678.github.io/research/software/kinmix.html







Thanks! Questions?